Les liens provenant de pages apprentées sont-ils vraiment sous-pondérés par Google ?

Retour au sommaire de la lettre

Domaine :	Recherche	**Référencement**
Niveau :	Pour tous	**Avancé**

Depuis de nombreuses années que sont explorés les algorithmes de pertinence des moteurs de recherche, il semblerait qu'un critère important, en termes de netlinking, soit la nécessité d'obtenir des liens "non apparentés" : sites différents, auteurs distincts, externes plutôt qu'internes, etc. Mais ce point n'était pas si simple à prouver que cela. Or, un brevet déposé par Google semble nous donner quelques indications complémentaires à ce sujet. La lecture approfondie de ce brevet semble indiquer, s'il est mis en place par le moteur de recherche, qu'effectivement, tous les liens ne sont pas égaux devant le Dieu Google...

En parcourant les blogs et les sites sur le référencement, un conseil revient régulièrement : pour améliorer son netlinking pour Google, il vaudrait mieux privilégier les liens émanant de noms de domaines différents (en d'autres termes, un lien externe a "plus de poids" qu'un lien interne).

Et ces conseils sont parfois assortis d'une autre conjecture : recevoir une multitude de liens émanant d'un seul et même site n'est pas optimal, car tous les liens ne sont pas pris en compte ou (cela dépend des auteurs) n'ont pas le même poids dans le calcul des scores des pages...

Ces "bonnes pratiques" sont tirées d'observations et de tests empiriques, mais force est de constater qu'il n'existe pas de démonstrations étayées par des études statistiques ou des preuves convaincantes de leur réelle efficacité.

Un brevet récemment publié par Google (Mai 2010) donne néanmoins un peu plus de crédit à ces affirmations. Intitulé "***Determining quality of linked documents***" (détermination de la qualité de documents liés), ce brevet décrit en fait une méthode d'évaluation de la qualité des liens dont les effets pourraient justifier les conseils cités plus haut.

Nous allons d'abord essayer de comprendre pourquoi il peut être utile de donner plus de poids à certains liens qu'à d'autres, avant de s'intéresser à la méthode décrite dans le brevet.

Dans la formule initiale du PageRank, tous les liens sont égaux

Dans la formule du PageRank publiée dans l'article de Larry Page et Serguey Brin (les fondateurs de Google), tous les liens sont pris en compte dans le calcul du score. Bien sûr, le PageRank des pages est différent, et le "jus" transmis d'une page à une autre dépend du PageRank de la page de départ et du nombre de liens, mais la formule est la même quel que soit la nature du lien. Par exemple, aucune différence n'est faite entre les liens internes (provenant du même domaine ou du même sous domaine) ou externes. On ne tient pas compte non plus de l'emplacement du lien dans la page.

Cette approche simple (d'aucuns diront simpliste), a été critiquée par certains chercheurs qui y voient un défaut majeur de l'algorithme : ce score, censé représenter l'importance d'une page sur la Toile, repose sur un modèle déconnecté de la réalité, à savoir le surfeur aléatoire.

Du surfeur aléatoire au surfeur intelligent, intentionnel ou rationnel

En effet, le score du PageRank représente en théorie la probabilité pour qu'un individu, qui clique au hasard sur les liens, finisse par atteindre une page donnée. Dans ce modèle (baptisé "surfeur aléatoire"), l'internaute peut aussi de temps en temps se lasser et entrer une adresse au hasard dans la barre de recherche du navigateur. Il se "téléporte" ainsi d'une zone d'Internet à une autre. Ces cas de "téléportation" correspondent mathématiquement dans la formule au facteur d'atténuation (damping factor).

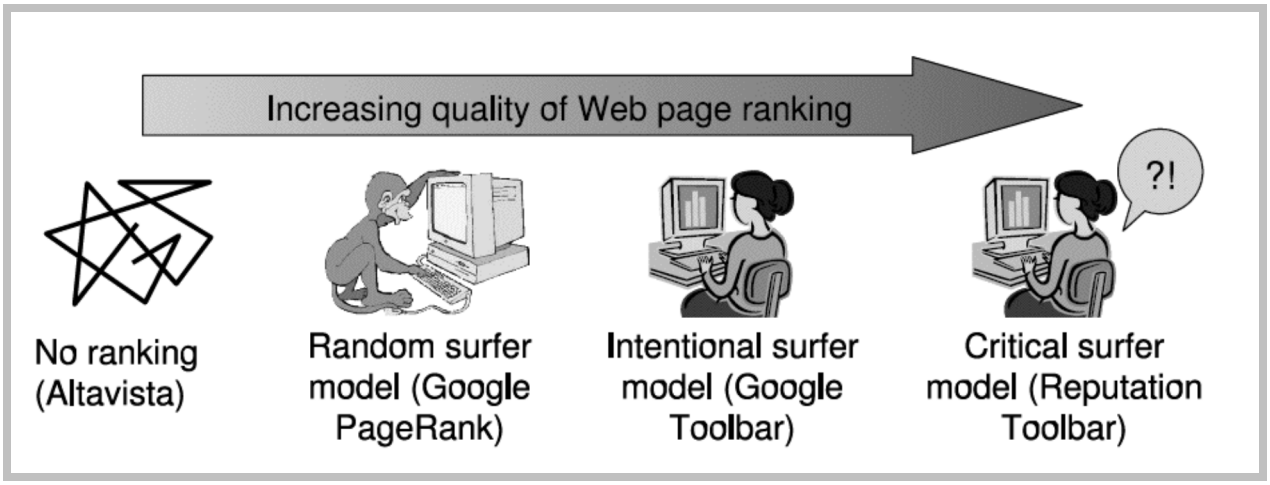
On comprend que ce modèle peut sembler naïf et assez éloigné du comportement réel d'un internaute.

Déjà, un internaute ne clique pas au hasard : il choisit ses liens en fonction d'un objectif : chercher une information, écouter un mp3, consulter sa messagerie...

Plusieurs approches ont été proposées pour améliorer le modèle du surfeur aléatoire :

Le surfeur intentionnel

Dans cette approche (décrite par Audun Jøsang), l'importance des pages prend en compte le nombre de visites reçues par des visiteurs réels sur les pages. Les données nécessaires sont censées être fournies par une toolbar (comme la Google toolbar). Le score final est une combinaison du résultat du calcul d'un PageRank normal pondéré par les résultats de l'algorithme basé sur le surfeur intentionnel.



Ce schéma illustre la vision d'Audun Jøsang sur l'évolution des modèles théoriques pour classer les pages. Après le surfeur intentionnel, il prédit l'avènement d'un modèle baptisé "surfeur critique" qui tient compte non seulement du trafic mais aussi de scores de qualité recueillis auprès des internautes.

Le surfeur intelligent

Quelques chercheurs de l'université de Washington ont proposé en 2002 le modèle du "surfeur intelligent". L'idée est de calculer un score qui tient compte des intentions de l'internaute lorsqu'il surfe sur le web. Il part du principe aussi qu'un score utilisé pour classer des pages dans un moteur de recherche doit tenir compte aussi du comportement des vrais utilisateurs du moteur.
Chaque lien est évalué en fonction de son intérêt par rapport aux termes d'une requête donnée. Le calcul du PageRank se fait tout simplement entre les seules pages qui contiennent le terme de la requête...

Et... la version de Google : le surfeur raisonnable !

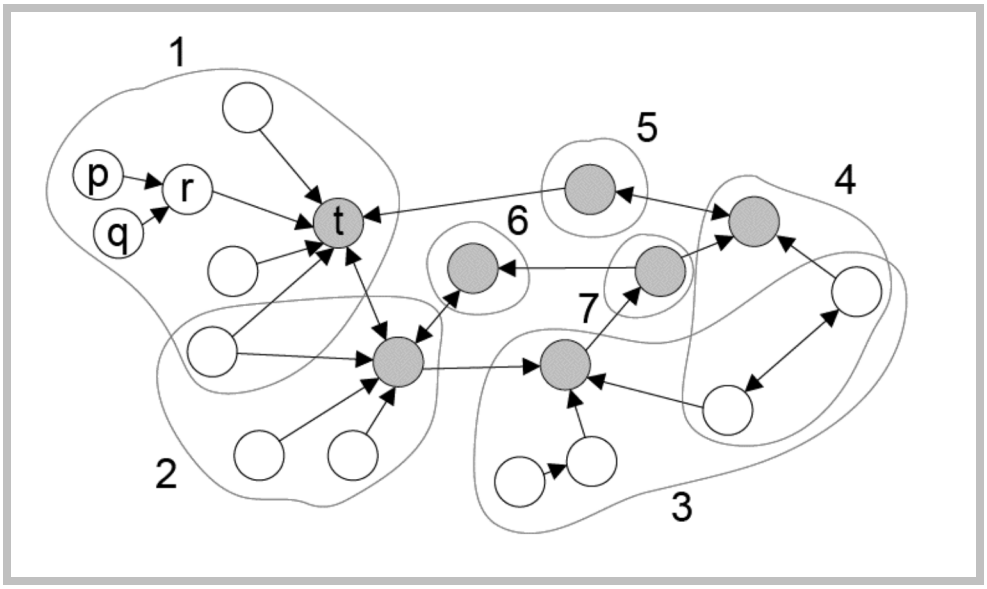
Les équipes de Google ont elles même inventé un modèle amélioré du surfer aléatoire : le surfeur raisonnable (*reasonable surfer*), décrit dans un brevet publié en mai 2010 (mais déposé en 2004, d'ailleurs en même temps que le brevet décrit dans cet article, ce qui évidemment suscite la curiosité tant ces deux approches sont complémentaires...).

L'idée ici aussi est de constater que quand un internaute consulte une page contenant des liens, la probabilité qu'un internaute clique sur un lien est plus grande pour certains liens que pour d'autres. Le brevet cite en particulier des exemples de liens peu susceptibles d'être cliqué souvent : les pages de "Conditions générales", les liens à caractère publicitaire, ou les liens sans rapport avec le document.

Il faut aussi compter avec le linkspam

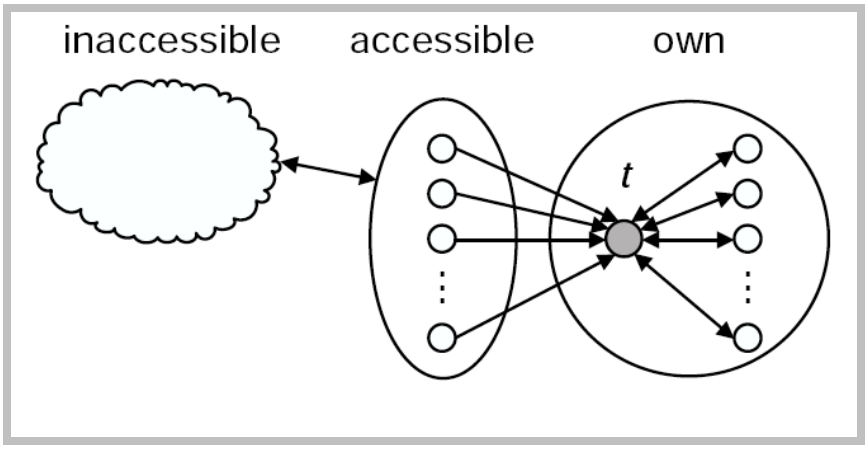
Déjà, sur un web "idéal" dénué de spam, considérer tous les liens comme identiques semble poser des problèmes théoriques et pratiques. Mais de plus, les webmasters ont appris à composer avec l'algorithme du PageRank et élaborer des stratégies efficaces pour "doper" leurs sites à l'aide d'achet de PageRank, d'échanges de liens, de fermes de liens, de création de galaxie de sites etc...

En effet, on peut démontrer mathématiquement qu'il suffit de créer des millions de pages nouvelles, et de les relier à une page que l'on souhaite "doper" pour lui donner un PageRank déjà suffisant pour se positionner sur des requêtes concurrentielles.



Un schéma de "ferme de pages" utilisé dans une stratégie de linkspam décrite dans un article d'Hector Garcia Molina et Zoltan Gyongyi (Link Spam Alliances). L'objectif est de "doper" la page t. Le résultat est obtenu en créant des liens depuis plusieurs galaxies de sites (numérotées 1, 2, 3, 4, 5, 6, 7). La structure de liens entre ces sites est complexe, irrégulière, et rend la détection de cette "ferme de spam" malaisée, ce qui est l'objectif de ses créateurs.

Dans la pratique, les référenceurs black hat utilisent en réalité une stratégie à plusieurs étages. Ils essaient de créer un volume de pages important, sur des noms de domaines qu'ils possèdent, et créent un maillage destiné à promouvoir une page donnée. Ensuite, ils essaient de créer des liens sur des sites qu'ils contrôlent directement (en créant automatiquement un grand nombre d'annuaires, de blogs, de *digg likes*) ou indirectement (par exemple en déposant des commentaires ou des posts sur des blogs ou des forums, ou, méthode la plus efficace, en hackant les sites pour en changer le contenu).



Stratégie de linkspam classique : utiliser ses propres sites/domaines d'abord (zones "owned"), ensuite essayer de créer des liens sur des sites que l'on contrôle (éventuellement en les hackant : zone accessible). La seule zone qui n'est affectée par le linkspam est dite "inaccessible". Le problème majeur étant évidemment de déterminer quelles pages sont dans les zones "owned" et "accessibles" pour un document donné...

Compte tenu de l'existence de ces pratiques de linkspam, il est clair qu'il est important pour un moteur comme Google de détecter ces pages, et éventuellement, de ne pas tenir compte des liens issus de telles tentatives de manipulation de son algorithme, ou, *a minima*, de déclasser ces liens.

Le brevet de Google : déterminer la qualité des documents en se servant des liens

Compte tenu des limites évoquées plus haut dans l'exploitation du PageRank comme principal score d'importance pour les pages, il serait assez logique que Google essaie d'améliorer son algorithme en prenant en compte des critères plus "qualitatifs" pour les liens.

C'est l'objectif du dispositif breveté par Google. Comme d'habitude avec les brevets américains, il faut éviter d'en conclure que Google utilise cette technique. En effet, il est possible aux USA de breveter des concepts, même assez vagues, et en particulier des algorithmes, alors qu'ils seraient refusés par un bureau des brevets européen. En conséquence, les entreprises américaines déposent beaucoup de brevets "par principe" afin de prouver ultérieurement une antériorité sur une technique ou une idée, ou pour gêner leurs concurrents qui voudraient explorer cette piste.

Ce brevet est néanmoins intéressant compte tenu des "auteurs" mentionnés :

- ***Amit Singhal***, l'un des rares "fellow engineers" distingué par Google, qui est responsable du département "*Search Quality*", et qui a ce titre est chargé d'améliorer en permanence la pertinence du moteur (l'*update* MayDay, selon Matt Cutts, est un pur produit du travail de l'équipe d'Amit Singhal).

- ***Krishna Bharat*** ensuite, le "père" de Google News, est l'auteur de l'algorithme "Hilltop", pour lequel il a développé le Localrank : on reconnait d'ailleurs quelques concepts communs au Localrank et à Hilltop dans ce brevet. Amit Singhal et Krishna Bharat sont les piliers historiques de l'algorithme de Google, une entreprise qu'ils ont rejoint très tôt (Krishna Bharat en 1999, Amit Singhal en 2000).

- Le dernier auteur, ***Paul Haahr***, est un des membres de l'équipe d'Amit Singhal et travaille chez Google depuis 2002. Paul Haahr est un développeur émérite, et a été la cheville ouvrière de nombreux changements majeurs de l'algorithme, et on retrouve sa "patte" dans, par exemple :
* la prise en compte intelligente des *stop words* (mots vides) ;
* l'introduction de certains raffinements de requêtes ;
* la détection des pages faisant autorité sur certaines requêtes ;
* la prise en compte de la "temporalité" dans les scores utilisés par Google.

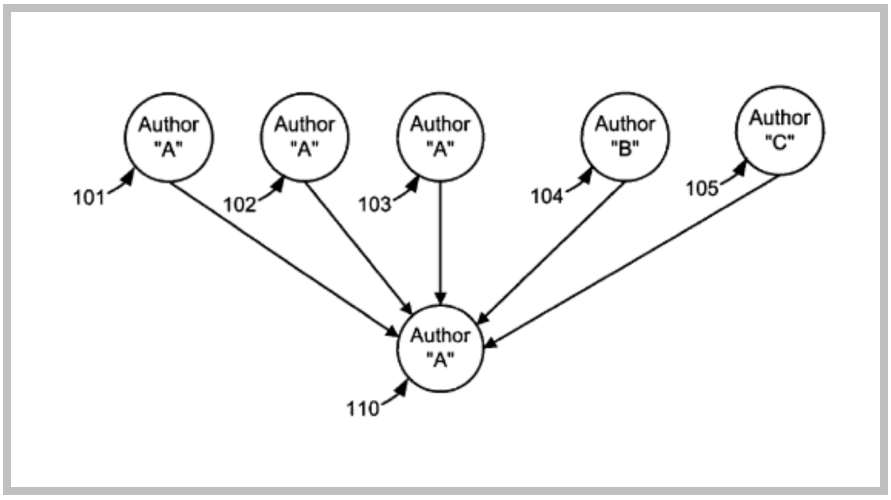
Le principe de l'approche

Le brevet décrit une méthode pour calculer un "score" permettant de classer des documents liés entre eux. Ici, l'objectif est de déterminer le degré d'"affiliation", terme anglais qu'il faudrait traduire ici par "degré de parenté entre les documents".

L'objectif est également de produire un score permettant de classer les documents en fonction de leur "qualité". Ce score tient compte des liens reliant les documents, comme le PageRank, mais il faut noter que le brevet ne cite jamais le "PageRank", ce qui est assez étonnant, puisque l'on devine que ce score peut servir, soit à l'intérieur du calcul du PageRank, soit dans le calcul d'une note complémentaire au PageRank classique.

Le problème des pages "apparentées"

L'objectif de la technique brevetée par Google est de résoudre le problème de la prise en compte des liens dans une situation comme celle décrite dans le schéma ci-dessous.



On voit que trois pages émanant de l'auteur A pointent vers le document 110 (qui est également produit par l'auteur A). Et deux autres liens émanant de deux autres auteurs différents pointent vers le même document 110.

Si tous les liens sont pris en compte de la même façon, 60% du score de la page 110 (3/5e), produite par l'auteur A, dépend de pages produites par ce même auteur. Si l'on compte les "votes uniques", le lien émanant de 104 ou de 105 ne peut pas compter trois fois moins que le vote émanant de l'auteur de la page !

On peut raisonner aussi en terme de qualité : comme la page 101 et la page 110 sont apparentées, on peut estimer que ces deux pages doivent avoir des scores de qualités similaires => les liens en provenance de pages apparentées ne peuvent pas augmenter les scores d'une page au-delà d'une certaine limite. Dans ce cas précis, ce n'est pas parce que 110 reçoit des liens de 101,102,103 que la "qualité" du document a été multipliée par trois !

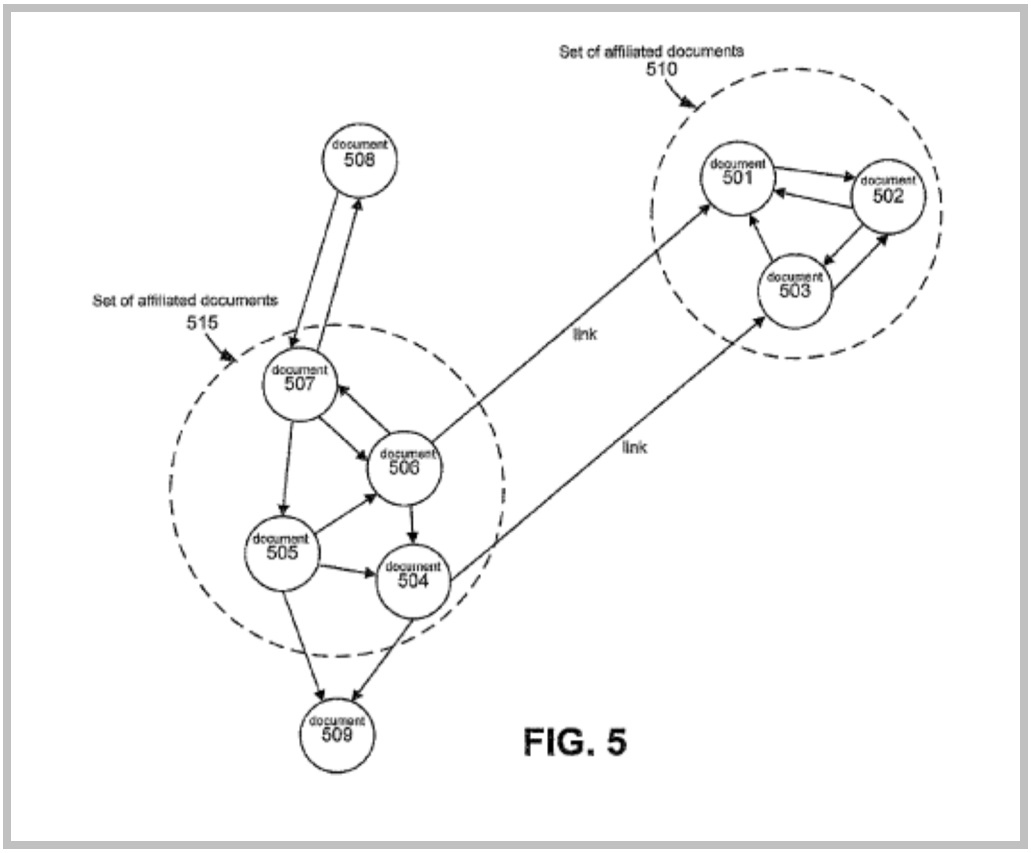
Comment déterminer les pages qui sont "apparentées"

Evidemment, comme le score dépend de l'identification de pages "apparentées", le système doit commencer par identifier ces pages "cousines". L'idée ici est avant tout de repérer que certaines pages ont les mêmes auteurs, appartiennent à la même entité, reprennent les mêmes contenus, ou sont produites par un groupe de personnes suffisamment proches pour que l'on considère qu'elles représentent un seul et même vote.

Le brevet présente plusieurs critères ou signaux qu'il est possible d'exploiter pour reconnaître que des pages doivent être considérées comme ayant un auteur unique :

- **l'analyse des graphes de liens** : on pense évidemment ici à un algorithme similaire à celui décrit dans cet article : Trawling the Web for emerging cyber-communities (http://www8.org/w8-papers/4a-search-mining/trawling/trawling.html). Logiquement, l'analyse des liens doit permettre de déterminer facilement les pages potentiellement apparentées.

- **l'analyse du maillage**, non seulement des pages entre elles, mais des sites ou des domaines qui hébergent les pages (l'analyse au niveau d'un site permet de détecter des relations qui ne se voient pas en analysant le maillage au niveau des pages).



L'analyse du maillage de liens et d'autres critères permet de regrouper les documents en groupes de pages apparentées disjoints, et de déterminer que, par exemple, les documents 501 et 506, bien que liés entre eux, ne sont pas apparentés, alors que les documents 505 et 506 sont apparentés. Pourtant, ces documents sont reliés entre eux de manière identique l'analyse du trafic vers les documents : les pages que de nombreux utilisateurs visitent au cours de la même session sont susceptibles d'être apparentées.

- **nom d'hôtes similaire** (noms de domaines, ou sous domaines).

- **IP similaires** : les documents sont hébergés au même endroit, et partagent une ip dont les deux ou trois premiers octets sont identiques.

Le système ne cherche pas à déterminer un score de degré de parenté, mais plutôt une réponse binaire : les documents sont soit considérés comme "apparentés", ou comme "non apparentés". En fonction de ce critère, le lien aura un poids différent dans l'algorithme.

Les deux systèmes de poids :

- Pour les backlinks provenant de pages "**non apparentées**", le "jus de lien" transmis à une page reste la somme des "jus" transmis par chaque backlink.

Pour les backlinks provenant de pages "**apparentées**" : une fonction de type "maximum" vient rogner la quantité de jus transmise à la page. On ne connait pas ce "maximum", donc on ne sait pas à partir de quel niveau les liens supplémentaires émanant de pages "apparentées" ne comptent plus dans l'algorithme.

Quelles sont les conséquences pour le référencement ?

Si cette technique est réellement utilisée, les conséquences pour les stratégies de netlinking sont importantes, même si ce qui suit fait souvent partie des "bonnes pratiques" popularisées par de nombreux experts en référencement.

Il faut tenir compte de la diversité des domaines, et des sources de liens, pour évaluer la qualité d'un netlinking. Compter juste le nombre de backlinks peut conduire à ne pas voir que 90% de ses liens provient d'un seul site. Or les 10 000 liens apportés par un seul site, ne comptent pas autant que 10 000 liens apportés par des sites différents !

Certains types d'échanges ne trompent pas un moteur de recherche comme Google : inutile de créer des échanges "triangulaires", ou entre pages différentes d'un même site. Ces schémas, s'ils sont reproduits de manière industrielle, sont détectés, et deviennent partiellement inefficaces.

Les liens internes, si la technique présentée dans ce brevet est appliquée, comptent moins dans l'algorithme que les liens émanant de sources externes et non apparentées.

Les stratégies à base de "galaxie de sites", de "fermes de spam" peuvent être totalement inefficaces, si ces pages sont visiblement "apparentées".

Par contre, rien n'indique dans ce brevet qu'une pénalité pourrait être infligée à un site qui reçoit des milliers de liens d'un autre site, parce qu'un lien partenaire a été placé dans le footer.

Le système "réduit" l'impact de ces liens, mais ne les élimine pas, et ce serait une erreur de demander à vos partenaires de retirer ce genre de liens sous prétexte que toutes ces pages sont apparentées...

Rien de neuf sous le soleil, mais une tendance se dégage...

On a vu que la publication de ce brevet ne révèle pas des changements sensationnels, mais il vient confirmer une source possible pour un comportement observé empiriquement par certains experts en référencement.

C'est aussi un indice de plus qu'une évolution s'est produite depuis quelques années dans les algorithmes des moteurs de recherche, et que l'implémentation naïve de l'algorithme du PageRank n'est plus d'actualité. Le modèle du surfeur "aléatoire" est dépassé, et quel que soit le nom qu'on lui donne ("intelligent", "rationnel", "critique", "raisonnable", ou "intentionnel"), la prise en compte du comportement réel de l'utilisateur est à présent au cœur des algorithmes de classement des moteurs. Ce brevet s'attaque à un aspect précis du comportement réel des utilisateurs : certains maitrisent de nombreuses pages du web, pour des raisons légitimes ou non, et il faut en tenir compte dans le classement des pages.

Toutes ces évolutions vers des algorithmes plus sophistiqués ont une conséquence : tous les liens ne comptent plus de la même façon dans l'algorithme, et ce, à plusieurs endroits de la fonction d'évaluation complexe qui sert à classer les pages dans un moteur de recherche comme Google. A vous d'en tirer les conséquences dans vos stratégies de référencement...

BIBLIOGRAPHIE

Trust and Reputation Systems
Audun Jøsang QUT, Brisbane, Australia
http://persons.unik.no/josang/papers/Jos2007-FOSAD.pdf

The Intelligent Surfer: Probabilistic Combination of Link and Content Information in PageRank
Matthew Richardson Pedro Domingos - Department of Computer Science and Engineering
University of Washington
http://alchemy.cs.washington.edu/papers/pdfs/richardson-domingos02a.pdf

Who links to whom : Mining Linkage between Web Sites
Krishna Bharat, Bay-Wei Chang, Monika Henzinger, Mathias Ruhl
Google, MIT Cambridge USA
http://people.csail.mit.edu/ruhl/papers/2001-icdm.pdf

PageRank Increase under Different Collusion Topologies

Ricardo Baeza-Yates, Carlos Castillo, Vicente Lopez
ICREA Uiversité Pompeu Fabra, Université du Chili
http://ramsesii.upf.es:8080/alfa/research_reports/baeza_05_PageRank_increase_collusion.pdf

Link Spam Alliances
Zoltan Gyöngyi, Hector Garcia-Molina
Université de Stanford
http://infolab.stanford.edu/~zoltan/publications/gyongyi2005link.pdf

Web Spam Taxonomy
Zoltan Gyöngyi, Hector Garcia-Molina
Université de Stanford
http://airweb.cse.lehigh.edu/2005/gyongyi.pdf

A Cautious Surfer for PageRank
Lan Nie, Baoning Wu, Brian D. Davison
Department of Computer Science & Engineering Lehigh University Bethlehem
http://www2007.org/posters/poster1038.pdf

Link-Based Similarity Search to Fight Web Spam
Andras A. Benczur, Karoly Csalogany, Tamas Sarlos
Académie des Sciences Hongroise et Université Eotvos Budapest
http://airweb.cse.lehigh.edu/2006/benczur.pdf

Trawling the Web for emerging cyber-communities
Ravi Kumar, Prabhakar Raghavan, Sridhar Rajagopalan, Andrew Tomkins -IBM Almaden Research Center
http://www8.org/w8-papers/4a-search-mining/trawling/trawling.html

BREVETS

Ranking documents based on user behavior and/or feature data
Invented by Jeffrey A. Dean, Corin Anderson and Alexis Battle
Assigned to Google inc
http://patft.uspto.gov/netacgi/nph-Parser?Sect1=PTO2&Sect2=HITOFF&u=%2Fnetahtml%2FPTO%2Fsearch-adv.htm&r=1&p=1&f=G&l=50&d=PTXT&S1=7,716,225.PN.&OS=pn/7,716,225&RS=PN/7,716,225
United States Patent 7,716,225 - Granted May 11, 2010 - Filed: June 17, 2004

Determining quality of linked documents
US Patent 7,783,639
http://patft.uspto.gov/netacgi/nph-Parser?Sect1=PTO2&Sect2=HITOFF&u=%2Fnetahtml%2FPTO%2Fsearch-adv.htm&r=1&p=1&f=G&l=50&d=PTXT&S1=7,783,639.PN.&OS=pn/7,783,639&RS=PN/7,783,639
Invented by Krishna Bharat, Amit Singhal, and Paul Haahr - Assigned to Google
Granted August 24, 2010 - Filed June 30, 2004

Method for node ranking in a linked database
US Patent 7,058,628
http://patft.uspto.gov/netacgi/nph-Parser?Sect1=PTO1&Sect2=HITOFF&d=PALL&p=1&u=%2Fnetahtml%2FPTO%2Fsrchnum.htm&r=1&f=G&l=50&s1=7,058,628.PN.&OS=PN/7,058,628&RS=PN/7,058,628
Inventor: Lawrence Page
Assignee: The Board of Trustees of the Leland Stanford Junior University
Granted June 6, 2006
Filed July 2, 2001

***Philippe Yonnet,** Global SEO Strategist, WEB DMUK (Londres) – Easyroommate / Vivastreet*

Réagissez à cet article sur le blog des abonnés d'Abondance :
http://blog-abonnes.abondance.com/2010/11/les-liens-provenant-de-pages-apprentees.html